Jun 22nd, 3:30 PM - 4:30 PM

DataStaR: a data staging repository to support the sharing and publication of research data

Gail Steinhart
Cornell University, gss1@cornell.edu

DataStaR: A Data Staging Repository to Support the Sharing and Publication of Research Data

Gail Steinhart
Albert R. Mann Library, Cornell University, USA
GSS1@cornell.edu

Abstract

The opportunities for new discoveries made possible by the widespread availability of large amounts of data are much in the news, as are the challenges associated with sharing research data and preserving it for the long term. Cornell University Library is engaged in a pilot project (funded by the National Science Foundation) to test the feasibility of a local "staging" repository to support data sharing among research collaborators while research is in progress, and to provide tools and support to publish data to permanent disciplinary or institutional repositories. DataStaR (short for "Data Staging Repository") aims to leverage local support and infrastructure as well as external resources to ensure access to data. Researchers may store and share data with selected colleagues, select a disciplinary repository or Cornell's own institutional repository for data publication, create high quality metadata in the formats required by external repositories and Cornell's institutional repository, and obtain help from data librarians with any of these tasks. Supporting data sharing while research is in progress allows the library to engage with researchers much earlier in the research process, alerting DataStaR staff to opportunities to curate, publish and preserve research data. We describe the overall design and operation of the system, partnerships with Cornell researchers, and the benefits and challenges associated with taking this approach to data curation.

Keywords: data staging repository, data curation, data sharing, data publication

Introduction

"… more technical data have been collected in the past year alone than in all previous years since science began." – Alex Szalay, in the Wall Street Journal (Hotz, 2009)

That digital data are accumulating at an unprecedented rate is not really news anymore, although the issue continues to garner significant attention in popular publications (including The Wall Street Journal, The Economist, The Washington Post, Wired Magazine, and others in the past year alone), as well as scholarly publications (for example, special sections on "Big Data" and "Data Sharing" in Nature, in 2008 and 2009, respectively). The rapid growth of the volume of digital data presents both opportunities and challenges. As more data are collected and made available, opportunities arise for new kinds of integration and discovery that were previously impossible, yet the means for organizing, distributing, and preserving these data are imperfectly and unevenly developed across disciplines. There are multiple actors in this arena, new and established, including data centers, publishers, funders, commercial entities, and new types of collaborative efforts such as those funded by the National Science Foundation's (NSF) DataNet program [National Science Foundation Office of Cyberinfrastructure, 2007]. Research libraries may also have an important role to play, and this paper focuses on one of the efforts of Cornell University Library (CUL) to develop an effective strategy for dealing with research data created by Cornell researchers.

Data-related activities at CUL

Data curation and distribution is not an entirely new activity for CUL. Well-established data distribution activities at Cornell include the United States Department of Agriculture Economics, Statistics and Marketing Information System (USDA-ESMIS)[1], and the Cornell University Geospatial Information Repository (CUGIR)[2]. USDA-ESMIS is maintained by contract with the USDA, and contains thousands of reports and data sets from five USDA agencies. The materials cover agriculture and related topics in the United States and internationally. CUGIR is one of two statewide repositories in New York State for geographic information systems (GIS) data sets, containing over 7000 data sets and metadata on topics related to agriculture, ecology, natural resources, and human-environment interactions.

After the Atkins Report [Atkins et al., 2003], whose publication we might take as an approximate indicator of the start of the current and more intense period of interest in data-driven research and data curation, CUL began to consider its potential role in this arena more generally, forming the Data Working Group in 2006. The purpose of the group was to exchange information about CUL activities related to data curation, to review and exchange information about developments and activities in data curation more broadly, and to consider and recommend strategic opportunities for CUL to engage in the area of data curation. The group read widely - exploring developments in various disciplines, what other libraries were doing, and what was happening at Cornell University specifically (not just the library). The culmination of this work was a white paper that was shared with the entire CUL community [Steinhart et al., 2008]. The paper itself included selected environmental scans of relevant activities beyond (national and international activities, universities and academic libraries, and training and career activities) and within Cornell (within and outside of CUL), brief discussion of some of the major issues in data curation (financial sustainability, appraisal and selection, digital preservation, intellectual property, confidentiality and privacy, and participation), and concluded with a set of recommendations for action. Among the recommendations was one to develop local infrastructure in support of data curation, in contexts that make sense within the university. The DataStaR project is one such effort.

Rationale for a Data Staging Repository (DataStaR)

We had already met with some success assisting researchers in publishing data to existing data repositories [Steinhart & Lowe, 2007], and to Cornell's institutional repository when no suitable external repository existed, or as a back-up to copies deposited elsewhere. Briefly, library staff worked with a research group that was working to improve understanding of nutrient and sediment cycling in the Upper Susquehanna River basin. The group had an explicit interested in sharing their data and findings widely, because the Susquehanna is the largest tributary to the Chesapeake Bay, which has had a history of water quality problems – thus the work of this group had potential utility to managers, policy makers, and the general public. We helped the group select and use a domain-specific metadata standard and repository in order to document and publish their data sets. We also provided a training session on creating metadata so researchers would be able to do at least some of the work themselves. While these early efforts were much appreciated by the researchers we collaborated with, we realized that if we wanted to scale up and work with more researchers and research groups, our initial approach was not going to be sustainable. Librarians spent many hours working with researchers to understand, format, and document their data. To be able to engage more researchers in this activity without greatly increasing library staff time, we knew we would need researchers to be able to do much of the work of documenting data themselves, with tools that are easy to use and with assistance from librarians as needed. We had also received multiple requests for online storage to facilitate collaboration among researchers. The free options for sharing data online at Cornell generally have time or size limits; a secure 'dropbox' will retain material for up to two weeks, and the attachment size limit to Cornell's wiki installation is small (15MB; the dropbox also has a size limit). Other for-fee services are also available, but the relative lack of useful options for sharing data means Cornell researchers must look to departmental infrastructure or to other units for this type of support, or find a way to sustain it themselves.

These two needs (assistance with documenting and publishing data, and a means to share data with collaborators), suggested that some sort of collaborative space, with support for publishing "finished" data sets to permanent repositories, might be useful to Cornell researchers. We also felt that by having the opportunity to engage researchers early on in the process, when they ask for support for sharing data with colleagues, we could begin a conversation about sharing or publishing data more widely. A presentation and a later paper by Ann Green and Myron Gutmann [Green & Gutmann, 2007], where the authors describe opportunities for partnerships between institutional repositories and domain repositories prompted us to consider what a repository environment meeting the above needs might look like, and led to the submission of the grant proposal that funds the DataStaR project. Disciplinary repositories have some advantages over more generic repositories, including better support for specialized tools for discovering, accessing, and interacting with data, but the burden of supporting such functionality for multiple disciplines is probably more than local institutions can assume themselves. Yet disciplinary repositories can seem remote to researchers, and there is ample evidence to suggest that even when disciplinary repositories exist, researchers may fail to make use of them [e.g. Karasti et al., 2006; Costello, 2009]. This suggests a possible role for a local service provider to actively recruit data into disciplinary repositories, thus moving data from private workspaces to public spaces where it can be shared. Treloar et al. [2007] explored some similar ideas in describing a data curation continuum that spans boundaries between private workspaces, collaboration spaces, and publication spaces, and these concepts also inform our work.

Description of DataStaR

DataStaR is a platform, as well as a set of services provided by librarians, intended to support research data sharing and publication. We focus our efforts on "small science" data sets, our practical definition of "small" being those data sets that are small enough to download and use in their entirety, and that do not require specialized infrastructure to access. A researcher may create a new data set in DataStaR, with minimal metadata, and upload and associate one or more files with it. They may assign read-only or read and edit permissions to individuals or research groups, as necessary. At the time a data set is created, the researcher is asked to indicate a target date for publication or deaccessioning, with the default target date being one year from the date of creation. If the target date passes and the data set has not been published or removed, DataStaR administrators may contact the researcher to clarify whether the data set should be retained in DataStaR. These capabilities meet the basic requirements for sharing among researchers. Researchers wishing to use DataStaR as a platform to create metadata for publication to a supported external repository may select that repository at the time they create the data set, or at a later time. Selecting a repository for publication triggers the presentation of a metadata form for that repository, for completion by the researcher. Publication may be handled a variety of ways, depending on the characteristics of the selected repository.

The DataStaR application consists of a Fedora[3] repository for storage of data files, a semantic metadata store based on vitro, a semantic web application developed at Mann Library, and additional open-source components: DROID[4], for file format identification, and SWORD[5] for deposit to SWORD-compliant repositories), as well as custom code to manage access, and package data and metadata for download, etc. (Figure 1).

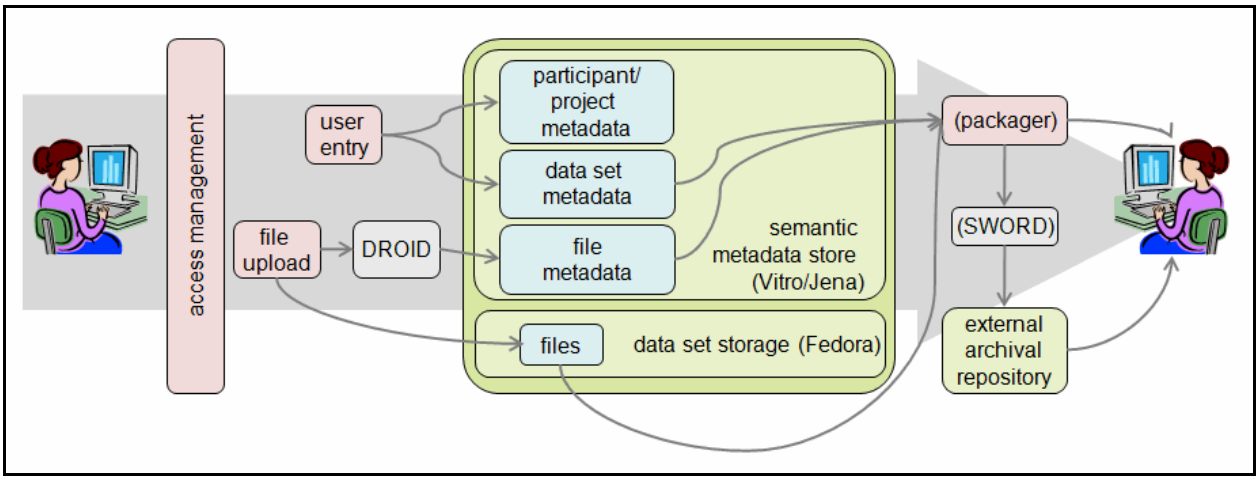

Figure 1. Overview of DataStaR architecture, from initial access by data set creator, left, to dissemination of data set to end user, right.

DataStaR itself requires very little metadata. Only four of the metadata elements require entry or selection on the part of the user: data set title, data set owner, metadata and data access permissions, and target repository for publication ("to be determined" is a valid selection). All other required elements are either supplied automatically, or have a default value if the user doesn't specify something different (Table 1). Additional, optional information may be provided, at the user's discretion.

Our approach to metadata management is described in more detail in [Lowe, 2009; Dietrich, accepted for publication], but merits some brief comments here. First, we adopt a semantic web approach to metadata on the assumption that scientific communities will increasingly adopt semantic web technologies, and that linked data will become increasingly common (although DataStaR, in its current form, does not yet support linked data), a development that will enhance interoperability and improve discovery across information resources. Semantic web efforts are already emerging in several research areas, and include the Marine Metadata Interoperability Project[6], an observational ontology for ecological data [Madin et al., 2007], and Bio2RDF, a semantic web portal for genomics data[7], to name just a few. Second, our treatment of metadata as a collection of statements, which can be reused by a researcher from one data set to the next, offers some potential efficiencies to DataStaR users as they create metadata for more data sets. It's not uncommon for a researcher or research group to make repeated use of the same methods or research sites, for example, and our approach makes that kind of reuse of information straightforward. In addition, we expect some users to deposit copies of data sets in both an appropriate disciplinary repository, as well as Cornell's institutional repository (eCommons@cornell)[8]. Mapping the relevant elements of a disciplinary metadata record to the fields required for eCommons is straightforward, thus saving users the added work of creating another metadata record for eCommons.

Table 1. Required and optional metadata.

Required metadata	Description
Owner	Data set owner. Supplied by user.
Originator	Person logged in at time data set is created. Supplied automatically.
Title	Title of data set. Supplied by user.
Identifier	Data set identifier. Supplied automatically.
Metadata and data access permissions	Read and/or write permissions for data and metadata. Supplied by user.
Publication repository	Target repository for publication ("to be determined" is allowable). Supplied by user.
Target date for publication (or deaccessioning)	Expected date data set will be published or deleted. Supplied by user, but if unspecified, the default is one year from date of creation.
File-level metadata	Person uploading file, file name, file format, file size, date and time of upload, file identifier, checksum. Supplied automatically.
Optional metadata	**Description**
Contact person	Contact person for data set. Supplied by user.
Relationship to research group	Relationship to DataStaR research group, if any. Supplied by user.
Relationship to other resources	Relationship to other resources listed in DataStaR. Supplied by user.
Citing publication	Publication citing data set. Supplied by user.
Abstract	Abstract. Supplied by user.
Temporal coverage	Temporal period covered by data set. Supplied by user.
Geographic coverage	Geographic coverage of data set. Supplied by user.
Usage rights	Information about rights asserted to resources, or conditions of use. Supplied by user.
Subject	Topic(s) covered by data set. Supplied by user.

Support for a discipline-specific metadata schema in DataStaR in achieved by first converting an XSD schema to an OWL ontology [Lowe, 2009]. The resulting OWL ontology is stored in the DataStaR vitro instance. Metadata forms for the schema can be generated semi-automatically, facilitating scalable support for multiple metadata specifications within DataStaR, but it remains a significant challenge to create usable forms with minimal manual work. At the point of publication or export, RDF is transformed into an XML document, which can then be validated against the original schema.

DataStaR operations and policies

One of the goals of DataStaR is to promote the publication of data so that it is preserved and accessible for the long term, yet DataStaR itself is a transitory environment. This concern for the long-term fate of data forced us to consider carefully our role and responsibilities with respect to data preservation, and how we should communicate those commitments to DataStaR users. To that end, we explored the Trustworthy Repositories Audit & Certification Criteria and Checklist (TRAC) [Center for Research Libraries & OCLC, 2007] as a framework for articulating system requirements and user and repository policies [Steinhart et al. 2009]. Briefly, we identified the TRAC requirements that we felt were applicable to the DataStaR environment, and settled on three mechanisms for demonstrating or documenting our approach to meeting those requirements: a set of repository policies, a deposit agreement, and system documentation (not yet publicly available)[9]. The repository policies describe the overall operation of DataStaR, including its mission statement, description of intended use, statements of intellectual property rights, collection development policy, data and metadata management policies and practices, digital preservation commitment, terms of use for depositors and users, and more. The depositor agreement includes a statement that the user agrees to abide by DataStaR's policies, and assigns sufficient non-exclusive rights to DataStaR for the repository to responsibly manage users' content. This agreement is largely based on the license agreement[10] in use by Cornell's institutional repository (eCommons@cornell), but was also influenced by CUGIR's data management policy[11], as well as the Data-PASS Deposit Agreement[12].

Current status

We have developed several partnerships with research teams that are eager to make use of DataStaR, including the following:

- Agriculture, Energy and the Environment Program (AEEP): A research group focused on understanding nutrient sources, sinks and management options in the Upper Susquehanna River basin, as well as the impact of hydraulic fracturing for natural gas development. To date, we have archived six data sets for this research group.
- Cayuga Lake Watershed Network: The Cayuga Lake Watershed Network is a community organization of citizens, businesses, associations, and local governments from throughout the Cayuga Lake Watershed: The Network is taking the lead on acquiring and documenting data sets related to the health of Cayuga Lake and its watershed. To date we've archived seven data sets for this group, with three more in process.
- Cornell Biological Field Station (CBFS): CBFS's mission is to provide facilities for long-term ecological research and to support the University's educational programs, with special emphasis on freshwater systems. The research program has produced a 50-year data set on the food web of Oneida Lake, New York, significant portions of which we've helped to archive.
- Cornell Plantations Natural Areas Program: The program manages 4300 acres of biologically diverse natural areas which are available for short and long-term research. Preliminary work on archiving selected data sets is underway.
- The Loon Project: The Loon Project is a long-term study focusing primarily on territoriality in loons. We've helped the project document and archive an extensive data set on loon

behavior on numerous lakes in northern Wisconsin, and anticipate documenting and archiving their extensive collection of audio recordings.
- The Virtual Center for Language Acquisition (VCLA): The VCLA examines language acquisition. Currently, we're working with the group to make audio recordings available for collaborative work.

To meet the needs of this diverse group of researchers, DataStaR will support publication to the repositories listed in Table 2.

Table 2. Target repositories for publication of data sets in DataStaR, and corresponding metadata specifications.

Repository	Metadata Specification
Cornell University Geospatial Information Repository (CUGIR, http://cugir.mannlib.cornell.edu/)	Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC-CSDGM, http://www.fgdc.gov/metadata/csdgm/)
eCommons@cornell (http://ecommons.cornell.edu/)	Modified Dublin Core
Knowledge Network for Biocomplexity (KNB, http://knb.ecoinformatics.org/)	Ecological Metadata Language (EML, http://knb.ecoinformatics.org/software/eml/)
Data Conservancy	TBD

Development of the DataStaR platform is not yet complete, and we continue to do the manual work of curating and publishing researchers' data sets. Nevertheless, even at this early stage, interest in DataStaR has spread, and the system is being adapted for use elsewhere. For example, the University of Melbourne, an early adopter of vitro for use as an expertise directory, is investigating their own version of DataStaR for Australian universities to register data sets with the developing Australian National Data Service (ANDS)[13]. We are also cooperating with the Data Conservancy[14], one of the first round of DataNet awards from the National Science Foundation, to explore the use of DataStaR as a "small science" front end to their data repository.

Challenges and Benefits to the DataStaR approach

We're optimistic about our approach to data curation in general, and our specific approach with the DataStaR repository in particular, but it does present some interesting challenges. We made the decision to adopt semantic web technology in order to facilitate extending the repository to support additional standards, allow reuse of pieces of metadata, as well as support future interoperability in a linked-data world. As noted earlier, development of semantic approaches to metadata would seem to signal that this is a sound decision. However, because we are bound to meet the requirements of the repositories to which DataStaR users wish to publish, and to date, their infrastructures still depend on XML schema-based metadata, DataStaR must support the production of XML metadata from RDF. We've been able to accomplish this, but it's perhaps not the most efficient means for producing XML metadata documents.

We've also encountered some more specific challenges with this approach. While we realize benefits in storing metadata as a collection of statements that can be reused, making changes to that core collection of statements can also introduce unintended consequences. For example, roles in an organization can be expected to change over time. It's appropriate to list the people and their roles at the time a data set is published, and not to update that information across all data sets

related to those same personnel with every subsequent change in staff or staff role. This problem has made it necessary for us to distinguish between and allow editing in both public and private graphs within DataStaR, to prevent changes to statements pertaining to one data set but not another from propagating across all data sets when that is not the desired behavior.

The extensibility of our approach and the fact that we've developed a reasonable method for embedding additional metadata specifications as OWL ontologies within DataStaR is an important advantage. However, focusing tool development on a single standard allows for greater effort and attention to usability and interface design issues, particularly for issues that are unique to a particular discipline or research community. In some cases, discipline-specific tools may offer a better user experience than DataStaR, at the expense of the flexibility of publishing to multiple repositories with less effort, or the ability to share preliminary data with only selected researchers.

One final challenge we anticipate is negotiating the varied and unique publication workflows of different repositories. For repositories with unique submission procedures or architecture, direct submission from DataStaR may not be possible. Publication in these cases may require the creation of a submission package in DataStaR that must then be submitted manually, either by a staff librarian, or the data owner. Nevertheless, we will be able to support submission to SWORD-compliant repositories, and some others.

In spite of the complexities we've outlined above, there are several important benefits we've realized in doing this work. Perhaps most importantly, we have gained (and continue to gain) valuable practical experience working with Cornell's researchers preparing, documenting, and publishing their data sets. The most tangible outcomes of this work are the more than twenty data sets that have been published, complete with high-quality metadata. Several more data sets are in preparation now. We've found that as researchers and librarians perform this work together, researchers learn more about ways to format, organize, and document their data that enhance reuse and its likely longevity, and the work of preparing data and metadata gets easier with each successive data set.

It remains to be seen whether the staging repository approach will prove the most effective way to promote effective data curation. The fact that DataStaR is a staging repository and not a permanent repository means that if CUL elects to pursue a different approach, DataStaR staff need only follow through on publication or deaccessioning of current content to meet the repository's obligations to its users, before transitioning to a new strategy. Nevertheless, as our work in this area has proceeded, we've observed an increased willingness to share data, and to engage the library earlier on in the research enterprise. Researchers who perhaps initially expressed some concerns about sharing data become more willing to do so, and spread the news among their colleagues that the library has been helpful to them in supporting their research in this way. Furthermore, researchers with whom we've already worked return to ask for input on data dissemination plans for grant proposals, in some cases, writing in partial salary support for librarians to assist with this work. We take this as a very strong vote of confidence in the library as it engages in this relatively new role, and are encouraged by the prospects this holds for the future.

Acknowledgements

The author gratefully acknowledges the contributions of the DataStaR team: Brian Caruso, Kathy Chiang, Jon Corson-Rikert, Dianne Dietrich, Ann Green, Huda Khan, Brian Lowe, Janet McCue, and Holly Mistlebauer. Dianne Dietrich and Huda Khan provided helpful comments on an earlier version of this paper.

This material is based upon work supported by the National Science Foundation under Grant No. III-0712989. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., & Messerschmitt, D. G. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue-ribbon advisory panel on cyberinfrastructur*e. Retrieved from: http://www.nsf.gov/od/oci/reports/atkins.pdf.

Center for Research Libraries, & OCLC. (2007). *Trustworthy repositories audit & certification (TRAC): Criteria and checklist*. Chicago; Dublin, Ohio: Center for Research Libraries; OCLC Online Computer Library Center, Inc. Retrieved from: http://bibpurl.oclc.org/web/16713.

Costello, M. (2009). Motivating online publication of data. *Bioscience, 59*(5), 418.

Dietrich, D. (accepted for publication). Metadata management in a data staging repository. *Journal of Library Metadata,*

Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services, 23*(1), 35-53.

Hotz, R. L. (2009, Aug 28). Currents -- science journal: A data deluge swamps science historians --- as paper trails fade, digital material grows in size and complexity; how to decipher those 80-column punch cards. *Wall Street Journal,* pp. A.6.

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work: CSCW: An International Journal, 15*(4), 321-358.

Lowe, B. (2009). DataStaR: Bridging XML and OWL in science metadata management. *Metadata and Semantic Research* 46, 141-150.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., & Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics, 2*(3), 279-296.

National Science Foundation Office of Cyberinfrastructure. (2007). *Sustainable digital data preservation and access network partners (DataNet).* Retrieved from: http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

Steinhart, G., Dietrich, D., & Green, A. (2009). Establishing trust in a chain of preservation: The TRAC checklist applied to a data staging repository (DataStaR). *D-Lib Magazine Magazine, 15*(9/10).

Steinhart, G. S., & Lowe, B. J. (2007). Data curation and distribution in support of Cornell University's upper Susquehanna agricultural ecology program. Paper presented at the Chapel Hill, NC. Retrieved from http://dspace.library.cornell.edu/handle/1813/7517

Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., et al. (2008). *Digital research data curation: Overview of issues, current activities, and opportunities for the Cornell university library* Retrieved from http://hdl.handle.net/1813/10903

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The data curation continuum: Managing data objects in institutional repositories. *D-Lib Magazine, 13*(9.)

Notes

[1] http://usda.mannlib.cornell.edu/
[2] http://cugir.mannlib.cornell.edu/
[3] http://fedoracommons.org/
[4] http://sourceforge.net/projects/droid/files/
[5] http://www.swordapp.org/
[6] http://marinemetadata.org/
[7] http://bio2rdf.org/
[8] http://ecommons.cornell.edu/
[9] http://datastar.mannlib.cornell.edu/index.jsp?primary=386684264
[10] http://ecommons.library.cornell.edu/rights.html#agreement
[11] http://cugir.mannlib.cornell.edu/CugirDataMgmtPolicy.20060828.pdf
[12] http://www.icpsr.umich.edu/files/DATAPASS/pdf/deposit-agreement.pdf
[13] http://ands.org.au/
[14] http://releases.jhu.edu/2009/10/02/sheridan-libraries-awarded-20-million-grant/